Scoring Errors on the WISC-III: A Study Across Levels of Education, Degree Fields, and Current Professional Positions

Elizabeth W. Brazelton, Robert Jackson, Joseph Buckhalt, Steve Shapiro, & Dianne Byrd
Auburn University

Abstract

Three WISC-III protocols and a questionnaire were sent to individuals representative of varying levels and fields of education and who held different professional positions for scoring. Data analyses of the 126 sets of protocols revealed that the Comprehension, Vocabulary, and Similarities subtests were scored incorrectly most often. Analyses also indicated that level of education was not related to scoring errors, nor was field of education, nor current professional position. There was a significant negative relationship between the number of WISC-III's administered and scoring errors with those respondents who had administered more than 100 WISC-III's having fewer errors than those who had administered 10 or less.

Elizabeth "Betty" Brazelton was a member of the faculty of Auburn University from 1989 until her death in 2002. Her final appointment was in the Department of Counseling and Counseling Psychology in the College of Education. This article represents two themes that permeated Betty's professional life as a school psychologist. First, as a researcher she was always a collaborator, frequently with colleagues, students and practitioners. She thrived in the social context of applied research. The second theme is exemplified by the central tenet of this article-that children deserve the highest quality services. Betty was a champion of all children and passed on that perspective to the many graduate students that she mentored into the profession. Her legacy will be sustained in all those who have been touched by her.

Previous research on individually administered intelligence tests has shown that numerous scoring errors occur on all forms and editions of the Wechsler Intelligence Scales (e.g.,Franklin, Stillman, Bureau, & Sabers, 1982; Gregory, 1999; Miller, Chansky, & Gredler, 1970; Slate & Jones, 1990; Slate, Jones, Covert, & Coulter, 1992; Written, Slate, Jones, Shine, & Raggio, 1994). Alfonso, Johnson, Patinella, and Rader(1998) provide a helpful table summarizing many of the published studies, the groups used, and the results. Many of these articles include suggestions to publishing companies to improve the scoring directions in the manuals and to professors to implement teaching techniques that would reduce or eliminate scoring errors (Alfonso & Pratt, 1997; Slate & Hunnicutt, 1988).

In fact, the manual for the Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991) discusses the extent to which scoring criteria were examined and changed during the standardization process of the new edition. Attention was focused on the Similarities, Vocabulary, and Comprehension subtests which were specifically mentioned as requiring the most scoring judgement. The manual reports interscorer reliability coefficients of .94 for Similarities, .92 for Vocabulary, and .90 for Comprehension. However, these reported reliability coefficients are based on the performance of scorers hired and trained by the Psychological Corporation for the specific purpose of standardizing the new edition. "The majority were certified or licensed professionals working in the school systems" (Wechsler, 1991, p. 27). They received written and oral feedback on their scoring and were required to submit a practice protocol prior to being approved for the position. The attention to interrater agreement in

the standardization process and the subsequent changes in the scoring criteria are welcome improvements; however, it is unknown whether those who are actually using the WISC-III are making scoring errors. Alfonso et al.(1998), using 15 graduate students in a school psychology program enrolled in their first individual intelligence testing course, found that out of 60 protocols, not one was scored correctly. As in other studies, the Comprehension and Similarities subtests had the most errors.

Additionally, a perusal of the literature regarding scoring errors on previous editions and forms of the Wechsler tests suggests that there have been few attempts to determine if any particular group made more errors than another group. Previous studies have generally used rather homogeneous samples in relation to the professional credentials and the level of training of the respondents. These research projects generally sampled either one profession, e.g. school psychologists, and only one or two levels of education, e.g. graduate students and those with master's degrees(see Alfonso & Pratt, 1997).Those studies that have compared groups have generally examined two groups, graduate students and professionals. Some of these studies have found no difference between the two groups, some have found that professionals made more errors than graduate students, and others found that graduate students made more errors than professionals (Kasper, Throne, & Schulman, 1968; Ryan, Prifitera, & Powers, 1983; Sherrets, Gard, & Langner, 1979; Slate, Jones, Murray, & Coulter, 1993). If training programs are to assume any responsibility for decreasing the number of errors in scoring, then the question of who is making the errors would seem an important one to address.

The current study was undertaken to determine the relationship between several variables and scoring errors. These variables were: experience in administering intelligence tests, number of WISC-III related workshops attended, the current career position, level of education, and field of highest degree held by respondents. Scoring errors were defined as errors in the assigning of credit to subtest items for all of the subtests with the exception of Block Design, Object Assembly, Digit Span, and Arithmetic. For these four subtests errors were defined as an incorrect raw score assigned to the subtest. Unlike some previous research projects, clerical errors were not considered scoring errors in this project as the focus was on errors that are more related to understanding the scoring criteria than on careless errors. For that reason we focused on the subtests that are the most likely to require judgement and have traditionally been found to require some subjective judgements i.e. Information, Picture Completion, Comprehension, Vocabulary, and Similarities.

Method

Participants

Recruitment of participants was carried out in the following ways: (1) letters were sent to professionals whose names were on lists that suggested they would be appropriate for the study(e.g, Alabama and Florida Associations of School Psychologists, Alabama Psychological Association whose members indicated that their practice included assessment of children); (2) a display was set up at the registration booth of the Georgia Association of School Psychologists' 1997 conference in Saint Simons Island, GA; (3)recruitment letters were sent to current graduate students, including those on internship, and recent graduates of the two departments at the university involved in the research project, including programs in school psychology, counseling psychology, counseling, clinical psychology, and experimental psychology; and (4) recruitment letters were sent to people whose names and addresses were provided by participants as colleagues who might be interested in participating. Approximately 400 recruitment letters were distributed, of which approximately 85% were to persons in the states of Alabama, Florida, and Georgia.

One hundred and ninety sets of materials were distributed and one hundred and twenty-six (66%) were returned during a 14 month period. Pertinent characteristics of the participants are presented in Tables 1 and 2. Sixty-eight percent were working in the schools as school psychologists, with forty-three percent having their highest

degree in school psychology. The largest percentage of the respondents (41%) had a master's degree plus additional hours of graduate work, followed by 22% with an Ed.S or equivalent degree, and 21% with a doctorate. All but one participant reported having had at least one formal course in individual intelligence testing. That person reported, spontaneously, that they had been trained by a doctoral level psychologist while on their pre-doctoral internship in counseling psychology. Additionally, eighty-one percent reported having had either a formal course specifically in the administration of the WISC-III (rather than other editions of the Wechsler Intelligence Scale for Children) or attending a workshop specifically related to the WISC-III. The number of workshops attended ranged from zero to four or more. While the field of highest degree varied, as well as the professional positions the respondents currently held, all acknowledged that they currently, or had previously, administered WISC-III's in their position or they were a graduate student who had administered WISC-III's as a function of coursework and/or in their practicum or internship.

Table 1
Current Position and Field of Highest Degree of Sample

Current Position	%	n	Field of Highest Degree	%	n
School Psychologist	68	81	School Psychology	43	51
			Counseling Psychology	06	07
Private Practice	16	19	Clinical Psychology	24	29
Student/Intern	12	14	Counseling	10	12
Other	04	05	Special Education	11	13
Other	06	07			

Table 2
Highest Degree of and Number of WISC-III's Administered by Sample

Highest Degree	%	n	WISC-III's Administered In Career	%	n
B.A.+ hrs.	03	03	0	1	1[a]
M.A.	14	16	1-10	6	7
M.A.+ hrs.	41	47	11-50	13	15
Ed.S	22	25	51-100	05	6
Doctorate	20	23	101-200	18	21
			200+	58	70

Note. [a] This category reflects a first year graduate student who had just completed an intellectual assessment class.

Materials

The protocols used in the study were actual protocols of public school students who had been referred for evaluation due to academic difficulties. The system school psychologist (Ed.D) selected three protocols and checked them for completeness and accuracy of administration. The answers were recopied onto fresh protocols to enhance the readability, omitting the scoring and all information from the protocol except for the student's birth date and the date of administration. Three doctoral-level psychologists scored the protocols independently: two university faculty members who teach intelligence testing and one practicing school psychologist. The senior author compared the scoring for consistency, and inconsistencies were reconciled by the senior author with the input of the expert scorers and those scorings became the "correct" scoring. The agreement among the expert scorers was high with disagreements on only six items out of the three protocols resulting in an interscorer agreement of 98.7%.

A questionnaire was also developed that requested information about the participants' education, experience, and current professional position. From this questionnaire the data used in the study as predictor variables was obtained.

Procedure

The recruitment letter explained the purpose of the project as well as the requirements for participation. In addition, the random selection of a participation incentive of $150 was described. A prepaid postcard was enclosed to be returned if the person was interested in participating. The postcard also had space for the name and address of a colleague whom they thought might be interested in participating. The recruitment letter was then sent to that person, also. Upon receipt of the postcard indicating a willingness to participate, each respondent was assigned a code number known only to a graduate assistant. This code number was placed on the three protocols and the questionnaire. These materials were sent to the participant along with an information letter, the directions for the study, and a self addressed, stamped envelope in which to return the materials. The participants were directed to score the three protocols in the same manner that they typically scored protocols, to fill out the questionnaire, and to return those materials in the addressed and stamped envelope. Upon receipt of the completed materials, the graduate assistant entered the code number in the pool from which the winner of the participation incentive of $150 would be drawn.

Each returned protocol was scored against the standard. The points awarded by the respondent for each individual item in each subtest was checked against the standard, and a point was assigned for every item on a subtest for which there was disagreement. If the scoring of the item was correct a value of 0 was assigned, representing no deviation from the correct score. The total number of errors per subtest for all three protocols was recorded. The Coding, Picture Arrangement, Arithmetic, and Block Design subtests were coded differently in that each of these subtest scores were compared with the standard. If there was a difference between the respondents' total raw score and the "correct" raw score, the difference was entered for each of the three protocols.

Results

Examining the results in total, no participant scored all three protocols perfectly. Comparison of subtests indicated that 98% (n = 124) of the participants made at least one error in scoring Comprehension, 96% made at least one error on the Vocabulary subtest, 75% on Similarities, 37% on Picture Completion, and 34% on Information. The total number of errors across all three protocols on Comprehension ranged from zero to eleven, inclusively (median = 3). The range of errors on Vocabulary was zero to eleven (median = 3), and the range of errors on Similarities was one to twelve (median = 1). The Information sub-test had a range of errors from zero to four (median = 0) and Picture Completion had a range of zero to eight (median = 0). On the Arithmetic subtest 35% of the respondents made at least one scoring error, on the Coding subtest 25% of the respondents made at least one error, 23% on Picture Arrange-

ment, 18% on Object Assembly, and 16% on Block Design.

Examination of the data led to the exclusion of six respondents from further analysis due to the extremely large number of scoring errors they made. The total number of errors by these respondents on the targeted subtests (Picture Completion, Information, Vocabulary, Similarities, Comprehension) were more than two and one-half standard deviations above the mean number of errors. The inclusion of these six cases made further analysis of central tendencies meaningless or misleading.

Looking at relationships between possible predictor variables (or independent variables), there were statistically significant relationships between the number of years administering individual intelligence tests and the number of WISC-III's administered in a career (r = .327, p < .01) as well as between the number of years administering individual intelligence test and the number of workshops related to the WISC-III's attended (r = .421, p < .01). There was also a statistically significant relationship between the number of WISC-III's administered in a career and the number of workshops attended (r = .481, p < .01).

The number of workshops attended varied by field of highest degree, current position, and level of education with field of highest degree having an $F(5,113) = 14.239$ ($p < .01$). Post hoc comparisons found that those with their highest degree in school psychology, counseling, special education, and other fields attended more workshops than did those in clinical psychology. An $F(4, 109) = 4.47$ ($p < .01$) was found for level of education and number of workshops attended with post hoc comparisons finding that those with a Master's or Ed.S degree attended more workshops than those with a bachelor's degree plus graduate hours, and those with an Ed.S or equivalent degree attended more than those with doctorates. Analysis of the number of workshops attended by current position found an $F(3, 115) = 40.410$ ($p < .01$). Post hoc comparisons found that school psychologists/psychometrists working in the schools attended more workshops than respondents in private practice or interns/practica students.

The relationship between number of scoring errors and experience was statistically insignificant (r = -.175) when experience was defined as number of years administering intelligence tests. When number of WISC-III's administered was used as the criterion of experience, there was a statistically significant $F(4,114) = 3.952$ ($p < .01$). Post hoc comparisons revealed that those respondents who had administered more than 100 WISC-III's had significantly fewer errors than those who had administered ten or fewer. Scoring errors were found to be not related to number of workshops attended, $F(4,115) = 1.849$, $p > .05$; level of education, $F(4,109) = .768$, $p > .05$; nor field of highest degree, $F(5, 113) = 2.224$, $p = .057$, although that F did approach significance. Scoring errors were statistically significantly related to the respondents current professional position with an $F(3,115) = 3.05$, $p < .05$. Post hoc comparisons found that those working in the schools as school psychometrists/psychologists had fewer errors than those who were in the "other" category.

In order to determine whether the relationship between the current professional position and number of errors was an accurate reflection of the data, an ANCOVA was conducted with the number of errors as the dependent variable, the current professional position as the predictor variable and the number of WISC-III's administered as the covariate. The resulting F test (see Table 3) was statistically insignificant with an $F(3,113) = 1.07$, $p > .05$.

Table 3
Analysis of Covariance of Number of Scoring Errors by Current Professional Position

Source	df	Mean Square	F	Eta Squared
Current Position	3	26.115	1.070	.028
Covariate	1	62.835	2.574	.022
Error	113	24.413		

Discussion

Some of the data reflect predictable relationships, such as the positive correlation between the number of years that respondents had administered intelligence tests, the number of WISC-III's administered, and the number of workshops they had attended. Also, it is probably not surprising that those whose highest degree was in school psychology administered more WISC-III's than those whose highest degree was in clinical psychology and that those whose highest degree was in clinical psychology administered fewer WISC-III's than all other groups. What was surprising to us was that those whose highest degree was in special education had administered more WISC-III's than those in school psychology. One possible explanation for the large number administered by those whose highest degree was in special education is that school districts may be hiring people who can do more than one job. A person with a master's degree in school psychology, for example, and a more advanced degree in special education might administer intelligence tests, academic tests, and teach special education. This is one area that needs to be further explored as it has implications for school psychologists and training programs.

That those working as school psychologists in the schools had administered more WISC-III's than practioners' in private practice who had administered more than practica/interns was also not surprising given the well documented data regarding the number of evaluations completed by school psychologists working in school systems (e.g. Goh, Teslow & Fuller, 1981; Stinnett, Havey & Oehler-Stinnett, 1994) and the fact that practica/interns haven't had the opportunity to administer many tests. It was also found that those whose highest degrees were in school psychology, counseling, special education, and other fields attended more workshops related to the WISC-III than did those whose highest degree was in clinical psychology; that school psychologists who worked in the school attended more workshops than those in private practice or interns/practica students; that those with Master's and Ed.S degrees attended more workshops than those with bachelor's degrees plus graduate hours, and that Ed.S level respondents attended more workshops than those with doctorates.

Other findings of this study supported the findings of Alfonso, Johnson, Patinella, and Rader (1998), who found that the 1991 version of the Wechsler Intelligence Scale for Children (Wechsler) is susceptible to scoring errors, just as previous research demonstrated the same finding for earlier versions of the Wechsler Intelligence Scales (see Alfonso & Pratt, 1997 for a summary of many research articles). Because of the difficulties and inconsistencies in determining scoring errors, it is impossible to determine if there were more or fewer errors between the WISC-R and the WISC-III for any group of participants, but it is important for all concerned to be aware that errors are still occurring. Generally, the more subjective subtests are still at greater risk for errors in scoring.

When analyzing the data for differences between groups as to the occurrence of scoring errors, it was determined that the relationship between experience and scoring errors was a

stronger relationships than was the field in which one had received their degree, level of education, or current professional position and scoring errors. That relationship suggests that the adage, "practice makes perfect," may, in fact, apply to scoring of the WISC-III.

While Kasper, Throne, and Schulman (1968) proposed that years of experience might solidify incorrect scoring patterns as those with more experience may rely more heavily on their experiences and memory rather than referring to the test manual, this study did not find that to be the case. When experience was defined as years administering intelligence tests, there was no statistically significant correlation between experience and number of scoring errors. However, when experience was defined as the number of WISC-III's administered, there was a statistically significant relationship with the number of scoring errors. Those who had administered more than 100 WISC-III's in their career had significantly fewer scoring errors than those who had administered between 1–10 WISC-III's.

Ryan, Prifitera, and Powers (1983) found no significant differences between Ph.Ds and graduate students on subtest or IQ scores, while Slate, Jones, Murray, and Coulter (1993) found practitioners more likely to make errors than graduate students. Our study found no differences between the various levels of education. This is a positive finding for those who train and those who hire school psychologists, as it suggests that there is no relationship between accuracy and an advanced degree. In some ways this is not surprising when one looks at programs of study for those programs that offer intelligence testing courses. Generally, a testing course is taught at the 6th year, or master's level, not at the doctoral level.

This study allows for some general, conservative conclusions and suggestions about error making with the WISC-III, but the sample was not a random sample of professionals who use the WISC-III. Of those professionals who indicated an interest in participating, more than 30% did not return the materials that were sent to them. Additionally, approximately 85% of the sample resided in Alabama, Florida, and Georgia. The impact of the monetary incentive on participation is not known. Consequently, the effect of the selection bias is unknown and is a confounding variable, as always. However, for this sample, the results are encouraging in that no identifiable group of professionals was found to be making more scoring errors than another group. With the knowledge of these findings we, as teachers of individual testing courses, know to encourage our students to get as much practice administering and scoring individual intelligence tests through practica and internship placements in order to increase their accuracy. It may also be that we need to consider increasing the number of individual administrations required in the testing courses that we teach.

References

Alfonso, V., Johnson, A., Patinella, L., & Rader, D. (1998). Common WISC-III examiner errors: Evidence from graduate students in training. *Psychology in the Schools*, 35, 119–125.

Alfonso, V., & Pratt, S. (1997). Issues and suggestions for training professionals in assessing intelligence. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 326–344). New York: Guilford.

Franklin, M., Stillman, P., Bureau, M., & Sabers, D. (1982). Examiner error in intelligence testing: Are you a source? *Psychology in the Schools*, 19, 563–569.

Goh, D., Teslow, C., & Fuller, G. (1981). The practice of psychological assessment among school psychologists. *Professional Psychology*, 12, 696–706.

Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice.* Boston: Allyn & Bacon.

Kasper, J., Throne, F., & Schulman, J. (1968). A study of the inter-judge reliability in scoring the responses of a group of mentally retarded boys to three WISC subscales. *Educational and Psychological Measurement, 28,* 469–477.

Miller, C., Chansky, N., & Gredler, G. (1970). Rater agreement on WISC protocols. *Psy-chology in the Schools, 7,* 190–193.

Ryan, J., Prifitera, A., & Powers, L. (1983). Scoring reliability on the WAIS-R. *Journal of Consulting and Clinical Psychology, 51,* 149–150.

Sherrets, S., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC pro-tocols. *Psychology in the Schools, 16,* 495–496.

Slate, J., & Hunnicut, L. (1988). Examiner errors on the Wechsler scales . *Journal of Psychoeducational Assessment, 6,* 280–288.

Slate, J., & Jones, C. (1990). Examiner errors on the WAIS-R: A source of concern. *The Journal of Psychology, 124,* 343–345.

Slate, J., Jones, C., Covert, T., & Coulter, C. (1992). Practioners' administration and scoring of the WISC-R: Evidence that we err. *Journal of School Psychology, 30,* 77–82.

Slate, J., Jones, C., Murray, R., & Coulter, C. (1993). Evidence that practioners err in administering and scoring the WAIS-R. *Measurement and Evaluation in Counseling and Development, 25,* 156–161.

Stinnett, T., Havey, J., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12,* 331–350.

Wechsler, David (1991). *Manual for the Wechsler Intelligence Scale for Children—Third Edition*. San Antonio, TX: Psychological Corporation.

Written, J., Slate, J., Jones, C., Shine, A., & Raggio, D. (1994). Examiner errors in administering and scoring the WPPSI-R. *Journal of Psychoeducational Assessment, 12,* 49–54.